AI 九思：用大语言模型焕新古汉语之美*

刘金柱[1] 王锦绣[1] 罗捷春[1] 李志芳[2,3] 袁方[1] 余静静[1] 龚丹[1] 谢雨霏[1] 罗婉滢[1] 郑苏楠[1] 陈旷心[1] 贺心雨[1] 张润哲[1] 夏婉婷[1] 谢佳延[1] 吕佳源[1] 吕萍[1] 余乐妍[1] 郑诗铭[1] 王金柳[1] 刘艺溶[4] 徐君词[1] 张雪晨[1] 冷谦益[5] 杨纯[6] 彭立雪[7] 张曼丽[8] 吴翊嘉[1] 李祎萌[1] 余锁湘[1] 汪靓[1] 刘根辉[1,9]

[1]（华中科技大学人文学院 武汉 430074）

[2]（武汉大学文学院 武汉 430062）

[3]（武汉大学古籍整理研究所 武汉 430062）

[4]（南京大学外国语学院 南京 210093）

[5]（科大讯飞核心研发平台 AI 资源部 合肥 230088）

[6]（成都市七中育才附属小学银杏校区 成都 610066）

[7]（武汉市第十二中学 武汉 430030）

[8]（安徽师范大学文学院 芜湖 241002）

[9]（华中科技大学筑牢中华民族共同体意识研究基地 武汉 430074）

摘要：[**目的/意义**]随着生成式人工智能（AIGC）的快速发展，大语言模型在通用领域展现出强大的语言理解和生成能力，但在古代汉语处理领域仍存在诸多局限。为应对这一挑战，华中科技大学研发了古汉语认知大语言模型“AI 九思”，旨在增强大语言模型在古汉语知识问答和理解应用方面的专业能力。[**方法/过程**]本文详细介绍了“AI 九思”的研发背景、数据集构建、模型训练过程及其在古汉语语言知识和语言能力方面的表现。[**结果/结论**]通过内测反馈，“AI 九思”在古汉语专业问答和理解应用任务上展现了显著优势，但也存在一些待改进之处。未来，团队计划进一步提升“AI 九思”的文本认知和多模态应用能力，优化用户交互体验，推动古汉语大语言模型向更高层次发展，促进古汉语研究向数智化阶段迈进。

关键词：AI 九思 古代汉语 数智化 大语言模型 多模态

分类号：G352

Abstract: [Purpose/Significance] With the rapid development of Generative Artificial Intelligence (AIGC), large language models have demonstrated powerful language understanding and generation capabilities in general domains. However, they still face many limitations in the field of ancient Chinese processing. To address this challenge, Huazhong University of Science and Technology has developed the

*本研究受到国家社科基金重大项目“明代至民国汉语非韵书罕见同音类聚文献的音韵研究及数据库建设”（21&ZD297）、国家社科基金重大项目“草创时期甲骨文考释文献的整理与研究”（20&ZD307）、全国高等院校古籍整理研究工作委员会资助项目“孙奭《册府元龟》音义辑考”（批准编号：1835）、中央高校基本科研业务费“《册府元龟》语料库建设、整理与研究”（2020WKYXZX004）、中央高校基本科研业务费“《册府元龟》引书研究”（21WKFZZX016）的资助。

本文的部分内容已刊发于全国古籍整理出版规划领导小组办公室编《古籍整理出版情况简报》，2024 年第 8 期（总第 642 期），第 27-29 页。此次刊发内容和观点有改动。

作者简介：刘金柱，博士研究生；王锦绣，博士研究生；罗捷春，硕士研究生；李志芳，博士研究生；袁方，硕士研究生；余静静，硕士研究生；龚丹，硕士研究生；谢雨霏，硕士研究生；罗婉滢，硕士研究生；郑苏楠，硕士研究生；陈旷心，硕士研究生；贺心雨，硕士研究生；张润哲，硕士研究生；夏婉婷，硕士研究生；谢佳延，硕士研究生；吕佳源，硕士研究生；吕萍，硕士研究生；余乐妍，硕士研究生；郑诗铭，硕士研究生；王金柳，硕士研究生；刘艺溶，本科生；徐君词，硕士研究生；张雪晨，硕士研究生；冷谦益，硕士，部门经理；杨纯，硕士，小学教师；彭立雪，硕士，中学教师；张曼丽，硕士研究生；吴翊嘉，本科生；李祎萌，本科生；余锁湘，本科生；汪靓，本科生；刘根辉，博士，教授，博士生导师，通信作者，E-mail：chnlab@mail.hust.edu.cn。

"AI Jiusi," a large language model for cognition of ancient Chinese, aiming to enhance the professional capabilities of LLM in knowledge question-answering and comprehension applications related to ancient Chinese. **[Method/Process]** This paper provides a detailed introduction to the research and development background, dataset construction, model training process, and performance in terms of ancient Chinese language knowledge and linguistic ability of "AI Jiusi." **[Results/Conclusions]** Based on internal testing feedback, "AI Jiusi" has shown significant advantages in professional question-answering and comprehension application tasks related to ancient Chinese, although there are areas that need improvement. In the future, the team plans to further enhance the text cognition and multimodal application capabilities of "AI Jiusi," optimize user interaction experience, and promote the development of LLMs for ancient Chinese to a higher level, facilitating the transition of ancient Chinese research into the digital and intelligent phase.
Keywords: AI Jiusi, Ancient Chinese, Digitalization and Intelligence, Large Language Model, Multimodal

1 引言

2022年被称为生成式人工智能（AIGC）元年，随后在不到两年的时间里，各种大模型如雨后春笋般地蜂拥而出，并在全球范围内演绎了一出AI领域“诸侯争霸”的壮阔画面。在通用领域，ChatGPT[1]、Llama[2]、Gemini[3]（Team G, Anil R, Borgeaud S, 2023）、文心一言[4]、通义千问[5]、讯飞星火[6]（科大讯飞，2023）等大语言模型得到广泛应用，展现出强大的语言理解和生成能力。然而，它们在处理古代汉语时的表现却不尽人意：在古代汉语信息处理下游任务上，如句读标点、文白翻译、实体识别等方面，受数据规模尤其是高质量数据规模的制约，通用大语言模型给出的结果往往并不能令人满意；在针对古代汉语的学习、研究、教学、应用等方面，面对专业用户提出的专业问题，通用大语言模型给出的答案看似文从字顺，实则错漏百出，难以满足专业需求。

为了应对通用大语言模型在专业领域应用方面的不足，国内多所高校几乎同时开始了面向古代汉语专业领域的大语言模型的研发探索，并相继取得了阶段性成果。北京师范大学于2023年11月开启古汉语文本理解的大语言模型“AI太炎1.0”的内测，并在2024年8月发布“AI太炎2.0”公众版[7]。南京农业大学于2023年12月发布古籍大语言模型“AI荀子1.0”，并于2024年5月发布“AI荀子2.0”[8]。华中科技大学于2024年元旦开启国内首个兼具古汉语知识问答和理解应用能力的古代汉语认知大语言模型“AI九思1.0”的内测，并由此开启了多模态古代汉语大语言模型的研发探索之路。

本文就“AI九思”的前世、今生与未来进行简要梳理和介绍，希望能在大数据、人工智能时代为古代汉语信息处理及数智化研究敞开一片全新的天地。

2 模型简介

“AI九思”是由华中科技大学人文学院汉籍数字化实验室、铸牢中华民族共同体意识研究基地刘根辉教授团队研发构建的一款既掌握古汉语专业知识、又具备古汉语理解应用能力的古汉语认知大语言模型。该模型具备较强的古汉语理解应用能力，能够完成智能句读标点、词法分析、文白翻译等古汉语信息

处理下游任务，同时掌握了文字、音韵、训诂、目录、版本、校勘等古汉语多领域的专业知识，能够为用户提供有关古代汉语知识的专业回答。

模型名称中的“九思”一词，出自《论语·季氏》：“君子有九思：视思明，听思聪，色思温，貌思恭，言思忠，事思敬，疑思问，忿思难，见得思义。”强调为人处世、一言一行都要认真思考和自我反省。做学问亦是如此。“AI 九思”从立项到上线，其间经历了研发方案的确定、数据的收集加工、模型的训练调优等，每一个环节都经过多次反复的深入思考、沟通和讨论，一步一个脚印地践行“君子九思”的精神，也自然铸就了“AI 九思”的精神内核。模型命名“九思”，同时也是为了致敬原华中工学院（华中科技大学前身）党委书记、院长、我国著名教育家朱九思先生。正是在朱先生的大力支持和推动下，才有了 1980 年成立的全国理工科大学中的第一个文科研究所——华中工学院中国语言研究所，这也标志着我国理工科高校的文科觉醒，并由此揭开了我国理工科高校人才培养模式转型的序幕：适应时代发展需求，从单科性人才培养向复合型、综合性人才培养模式转型。“AI 九思”也正是在学校文工交叉的学科发展优势下崭露头角的。

3 数据集构建

高质量的数据是大语言模型构建的基础。“AI 九思”的数据主要源于古代汉语领域 100 余种经典权威的书籍，其中也包括部分重要的古籍译注本、古代汉语经典教材、参考书等。在数据集构建过程中，团队严格按照科学化、规范化的流程，确保数据的准确性、完整性和实用性。团队首先对原始文献进行了数字化处理，通过扫描和 OCR 识别技术获取初步的电子文本；接着，按照不同任务类型和知识领域，制定了详细的标注规范，并依据规范分工合作，对这些初版电子文本进行人工筛选、标注、校对和编辑，确保了数据的准确性和完整性，从而形成了精校后的电子版本；最后，使用 Python 语言编程，按任务类型、知识领域，自动匹配预先设置的多样化 Prompt 指令模版，从而完成整个古代汉语大语言模型数据集的构建。

对于部分网络数据，我们仅采用北京师范大学开源的通假字资源库[9]和北京大学数字人文研究中心开源的、供科研使用的 GuNER2023 古籍命名实体识别数据集[10]。对于这部分网络数据，团队均进行了核验和二次加工处理，以确保其质量与团队自建数据的高度一致性。

最终，“AI 九思 1.0”古汉语认知大语言模型团队构建了涵盖古代汉语语言知识、古代汉语语言能力 2 大模块、11 个子类别、共计 11 万条的高质量古汉语大语言模型数据集。如图 1 所示。

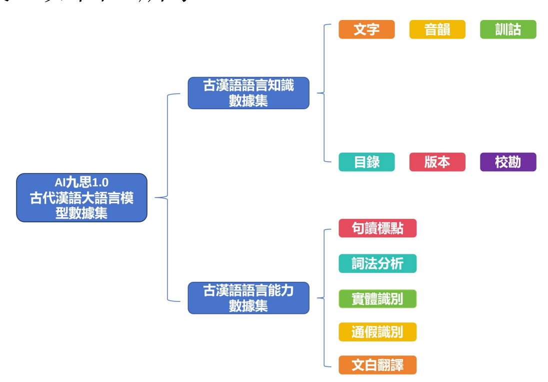


图 1 “AI 九思 1.0”数据集数据构成

4　模型训练

“AI 九思 1.0”以阿里云“通义千问-7B（Qwen-7B）”为基座模型，该模型是阿里云研发的通义千问大模型系列的 70 亿参数规模的模型。Qwen-7B 采用 Transformer 架构，并经过超过 2.4 万亿个 token 数据的预训练，使其在多个中英文下游评测任务上（涵盖常识推理、代码、数学、翻译等），效果显著超越同级别开源模型，甚至在部分指标上相比更大尺寸模型也有较强竞争力[5]。为了提升其在古汉语这一特定垂直领域的性能，团队使用了 4 张 A100-SXM4-80GB GPU 进行了深入的定制化训练，以增强大模型在古汉语这一垂直领域多项任务上、多类知识领域中的理解和生成能力。

在模型训练时，团队采取了多项优化措施以确保训练的高效性和有效性。我们先后进行了继续预训练（Continued Pretraining）和有监督微调（Supervised Fine-tuning）两个阶段的训练，并引入了 DeepSpeed ZeRO 2 技术，有效减少了显存的冗余占用，提高了训练过程中的资源利用率[11]。在继续预训练阶段，团队在自行构建的古汉语大语言模型数据集未标注版本上进行进一步的预训练。这一阶段的目标是让模型接触到丰富且多样的古汉语表达形式，从而更好地捕捉古汉语的语言结构和语义特征，增强其在处理古汉语文本时的泛化能力。在有监督微调阶段，团队设计了一系列针对古代汉语知识问答和理解应用的下游任务，如文字学知识问答、音韵学知识问答、训诂学知识问答、古汉语句读、古汉语翻译、古汉语实体识别等，并在自行构建的古汉语大语言模型数据集标注版本上进行了训练。在此过程中，我们采用了 LoRA（Low-Rank Adaptation）技术[12]，通过在模型中引入低秩矩阵来优化参数更新，不仅减少了所需的训练时间和计算成本，还有效防止了过拟合，显著提高了模型在目标任务上的表现。

通过上述训练步骤，“AI 九思 1.0”不仅继承了 Qwen-7B 的强大基础能力，还在古代汉语的知识问答和理解应用方面实现了显著的性能提升，为古代汉语这一特定领域的用户提供更加精准和丰富的交互体验。

5　模型的语言知识与语言能力

在古代汉语语言知识方面，“AI 九思 1.0”能够准确回答一部分有关古代汉语文字学、音韵学、训诂学等方面的专业问题，如“什么是异体字？”“什么是六书？”“什么是反切？”“什么是三十六字母？”“什么是训诂学？”“什么是形义统一？”等等；也能针对古典文献学领域关于目录学、版本学、校勘学等方面的一些专业问题给出满意的回答，如“什么是目录学？”“什么是刻本？”“什么是高丽本？”“校勘、理校、他校有什么异同？”等等。

在古代汉语语言能力方面，“AI 九思 1.0”可以快速、准确地完成古代汉语文本的断句和标点，较准确地完成古代汉语文本的自动分词和词性标注，自动进行古代汉语文本中的实体信息抽取，智能识别出古汉语文本中的通假字并进行注音、释义，并且能够为古诗词、文言文等古代汉语文本提供高质量的文白翻译。

6　模型内测反馈

2024 年 1 月 4 日—10 日，“AI 九思 1.0”开启了为期七天的内测。此次内测吸引了来自教育、科研和技术等不同领域的 237 位用户参与，覆盖了国内外 130 个不同的工作和学习单位，包括全国各地的高校、科研院所、中小学和企业，以及诸如英国伦敦大学、韩国庆星大学、泰国东南曼谷大学、拉脱维亚大

学等国外高校（如图 2）；其职业涵盖国内外高等院校的教师、研究员、AI 工程师，孔子学院（语合中心）外派国际中文教师及海外当地中文教师、小学至高中阶段语文教师、高校在读本硕博学生、语文编辑、古代汉语爱好者等等（如图 3）。其中高校教师占比接近 30%，高校学生占比超过半数。他们从各自的专业角度给“AI 九思 1.0”提出了宝贵的意见或建议。内测成员广泛的地域分布和多元化的成员结构，使得“AI 九思 1.0”得到了具有典型性和代表性的测试和评估。



图 2 参与内测单位分布图

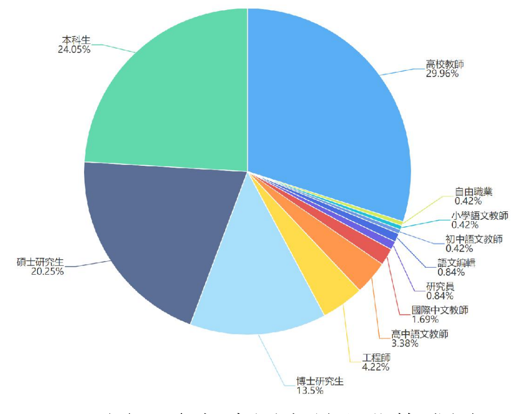


图 3 参与内测人员职业构成图

本次测评共收到反馈信息 400 余条，经过去重、整理和分类，得到 136 条有效反馈信息，其中既有鼓励性的用户评价，也包括专业性的用户意见和前瞻性的用户建议。绝大部分用户对“AI 九思 1.0”点赞并给予积极评价，认可其在古代汉语大语言模型研发领域的探索性和开拓性，认为“AI 九思 1.0”在古代汉语专业知识问答和古代汉语信息处理任务上具备了一定的能力。南京大学一位教授留言：“古汉语之美永远不会过时，华科大在汉语研究上常有独到的创见创新、开发和运用，鼓舞人心，引领潮流，致敬和感谢你们！”这是对本研发团队的莫大鼓励和鞭策。也有很多用户从各自的专业角度和使用体验出发，反馈了一些专业的意见，指出“AI 九思 1.0”在界面友好度、运行稳定性、数据丰富度、功能多样性、回答准确率等方面存在的不足。还有不少用户结合自己在古代汉语、人工智能、大语言模型、古代汉语信息处理等领域的研究与实践，提出了有关界面设计、数据建设、功能拓展、应用场景延伸、开放开源等方面的前瞻性建议。这些意见和建议为我们今后改进大语言模型提供了极有价值的参考。

7 模型迭代规划

“AI 九思 1.0”古代汉语大语言模型既是华中科技大学人文学院汉籍数字化实验室多年来深耕古代汉语信息处理相关学科交叉研究的薄发之作，也是研发团队探索和推动古代汉语大语言模型发展的第一步。目前，团队正致力于多模态古代汉语大语言模型“AI 九思 2.0”的研发，已取得阶段性突破性进展，并将在 2025 年 1 月下旬发布，2 月中下旬开启内测。

在新一代大语言模型中，我们着力推进以下几方面工作。

其一，文本认知能力提升。为有效提升“AI 九思”的文本理解与生成能力，有必要在已有的古代汉语文本数据集基础上，进一步扩充数据集。截至 2024 年底，研发团队已经完成十余万条数据集的精加工，其范围涵盖文字学、音韵学、训诂学、目录学、版本学、校勘学、方言学等古代汉语语言知识，也包括词法

分析、句读标点、文白翻译、命名实体识别、通假字识别、诗文典故识别等古代汉语语言理解及应用。

其二，图片理解能力扩展。早期版本的“AI 九思”，已具备了针对本文数据做出响应的单模态形式的大语言模型能力，随着大数据与人工智能技术的飞速发展，大语言模型必然朝着多模态形式的方向发展。但由于古代汉语处理对象的特殊性，现有的古代汉语大语言模型基本还不具备语音、图像的识别和加工能力。为此，本团队率先尝试扩展古代汉语大语言模型的图片理解能力，拟将“AI 九思 2.0”由基于文本单模态的语言模型，升级为能够同时处理文本和图片的多模态大语言模型。我们以甲金文字著录片为数据源，构建了一个包含 2000 多个字头，两万余张甲金文字图片的数据集。截至 2025 年 1 月中旬末，研发团队已经大致完成了文本模态和图片模态的有监督微调的工作，并将在 2025 年 1 月下旬发布“AI 九思 2.0”，随即开启内测。

其三，用户交互体验优化。针对“AI 九思 1.0”内测用户反馈的界面优化问题，我们对“AI 九思 2.0”的用户界面进行了全面升级，强化功能引导设计，提升用户操作的便捷性和直观性，确保用户在使用“AI 九思 2.0”过程中获得更加流畅和愉悦的交互体验。

8 结语

目前，“AI 九思 1.0”已经实现了从 0 到 1 的突破，研发团队将继续秉承和践行“君子九思”的精神，在各界的支持和鼓励下，兼顾质量和数量持续构建多元数据；挖掘用户专业需求，拓展延伸功能应用；细化完善界面设计；升级模型运行资源，提升模型运行稳定性。我们期待，古代汉语多模态大语言模型“AI 九思 2.0”的推出，能够不断推动古代汉语大语言模型向古代汉语大模型乃至古代汉语“智能体”的发展和转变，为探求古汉语之菁华、赓续古汉语之绝学、传承古汉语之美、继承和弘扬中华优秀传统文化做出应有的贡献！

参考文献：

[1] OpenAI. Introducing ChatGPT[EB/OL].https://openai.com/index/chatgpt/,2022-11-30.

[2] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprin t arXiv:2302.13971, 2023.

[3] Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models[J]. arXiv preprint a rXiv:2312.11805, 2023.

[4] Baidu. ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology [EB/OL]. https://research.baidu.com/Blog/index-view?id=183,2023-03-24.

[5] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[6] 科大讯飞. 科大讯飞发布星火认知大模型 [EB/OL].https://mp.weixin.qq.com/s/3esI9MJsHgHuMZHNOFuVuA,2023-05-07.

[7] 李绅,胡韧奋,王立军.古汉语大语言模型的构建及应用研究[J]. 语言战略研究, 2024,9(05):22-33.

[8] 王东波课题组. 荀子古籍大模型 [EB/OL].https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM,2024-11-30.

[9] 王兆基,张诗睿,胡韧奋,等. 古汉语通假字资源库的构建及应用研究[J]. 中文信息学报,2024,38(03):152-162.

[10] 北京大学人工智能研究院, 北京大学数字人文研究中心. CCL23 古籍命名实体识别评测 [EB/OL].https://guner2023.pkudh.org/,2023-04-01.

[11] Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.
[12] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
（通信作者：刘根辉 E-mail：chnlab@mail.hust.edu.cn）

作者贡献声明：

大语言模型“AI 九思”是在项目负责人刘根辉教授带领下的研发团队所有成员的集体智慧的结晶。

刘金柱：完成全部实验流程，论文初稿撰写；

王锦绣：负责数据集构建，论文修改；

罗捷春，李志芳，袁方，余静静，龚丹，谢雨霏，罗婉滢，郑苏楠，陈旷心，贺心雨，张润哲，夏婉婷，谢佳延，吕佳源，吕萍，余乐妍，郑诗铭，王金柳，刘艺溶，徐君词，张雪晨，冷谦益，杨纯，彭立雪，张曼丽，吴翊嘉，李祎萌，余锁湘，汪靓：数据集构建；

刘根辉：论文整体框架设计，论文定稿，研究思路制定。